ИНФОРМАТИКА, ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА И УПРАВЛЕНИЕ
INFORMATION TECHNOLOGY, COMPUTER SCIENCE, AND MANAGEMENT

UDC 004.11.6
https://doi.org/10.23947/2687-1653-2022-22-4-373-383

Original article

Determinants Factors in Predicting Life Expectancy Using Machine Learning

Brou Kouame Amos, Ivan Smirnov
Peoples' Friendship University of Russia (RUDN), 6, Miklikho-Maklaya St., Moscow, Russian Federation
✉ broukouameamos9@gmail.com

Abstract

Introduction. Life expectancy is, by definition, the average number of years a person can expect to live from birth to death. It is therefore the best indicator for assessing the health of human beings, but also a comprehensive index for assessing the level of economic development, education and health systems. From our extensive research, we have found that most existing studies contain qualitative analyses of one or a few factors. There is a lack of quantitative analyses of multiple factors, which leads to a situation where the predominant factor influencing life expectancy cannot be identified with precision. However, with the existence of various conditions and complications witnessed in society today, several factors need to be taken into consideration to predict life expectancy. Therefore, various machine learning models have been developed to predict life expectancy. The aim of this article is to identify the factors that determine life expectancy.

Materials and Methods. Our research uses the Pearson correlation coefficient to assess correlations between indicators, and we use multiple linear regression models, Ridge regression, and Lasso regression to measure the impact of each indicator on life expectancy. For model selection, the Akaike information criterion, the coefficient of variation and the mean square error were used. R^2 and the mean square error were used.

Results. Based on these criteria, multiple linear regression was selected for the development of the life expectancy prediction model, as this model obtained the smallest Akaike information criterion of 6109.07, an adjusted coefficient of 85 % and an RMSE of 3.85.

Conclusion and Discussion. At the end of our study, we concluded that the variables that best explain life expectancy are adult mortality, infant mortality, percentage of expenditure, measles, under-five mortality, polio, total expenditure, diphtheria, HIV/AIDS, GDP, longevity of 1.19 years, resource composition, and schooling.

The results of this analysis can be used by the World Health Organization and the health sectors to improve society.

Keywords: life expectancy, machine learning, machine learning models.

Acknowledgments. This paper has been supported by the RUDN University Strategic Academic Leadership Program.

For citation. Brou Kouame Amos, I. Smirnov. Determinants Factors in Predicting Life Expectancy Using Machine Learning. Advanced Engineering Research, 2022, vol. 22, no. 4, pp. 373–383. https://doi.org/10.23947/2687-1653-2022-22-4-373-383

Научная статья

Детерминирующие факторы в прогнозировании ожидаемой продолжительности жизни с помощью машинного обучения

Бру Куамэ Амос, И. Смирнов
Российский университет дружбы народов (РУДН), Российская Федерация, г. Москва, ул. Миклухо-Маклая, 6
✉ broukouameamos9@gmail.com

Аннотация

Введение. Ожидаемая продолжительность жизни — это, по определению, среднее количество лет, которое человек может прожить от рождения до смерти. Таким образом, это лучший индикатор для оценки здоровья

людей, а также комплексный индекс для оценки уровня экономического развития, систем образования и здравоохранения. В результате нашего обширного исследования мы обнаружили, что большинство существующих исследований содержат качественный анализ одного или нескольких факторов. Отсутствует количественный анализ множества факторов, что приводит к ситуации, когда невозможно точно определить преобладающий фактор, влияющий на продолжительность жизни. Однако при наличии различных состояний и осложнений, наблюдаемых сегодня в обществе, необходимо учитывать несколько факторов для прогнозирования ожидаемой продолжительности жизни. Поэтому были разработаны различные модели машинного обучения для прогнозирования продолжительности жизни.

Целью данной статьи является выявление факторов, определяющих продолжительность жизни.

Материалы и методы. В нашем исследовании используется коэффициент корреляции Пирсона для оценки корреляций между показателями, и мы используем несколько моделей линейной регрессии, регрессию Риджа и регрессию Лассо для измерения влияния каждого показателя на ожидаемую продолжительность жизни. Для выбора модели использовали информационный критерий Акаике, коэффициент вариации и среднеквадратичную ошибку. Использовались R^2 и среднеквадратическая ошибка.

Результаты исследования. На основании этих критериев для разработки модели прогнозирования ожидаемой продолжительности жизни была выбрана множественная линейная регрессия, поскольку эта модель получила наименьший информационный критерий Акаике 6109,07, скорректированный коэффициент 85 % и среднеквадратичное отклонение 3,85.

Обсуждение и заключения. В конце нашего исследования мы пришли к выводу, что переменными, которые лучше всего объясняют ожидаемую продолжительность жизни, являются взрослая смертность, младенческая смертность, процент расходов, корь, смертность детей в возрасте до пяти лет, полиомиелит, общие расходы, дифтерия, ВИЧ/СПИД, ВВП, продолжительность жизни 1,19 года, состав ресурсов и обучение.

Результаты этого анализа могут быть использованы Всемирной организацией здравоохранения и секторами здравоохранения для улучшения общества.

Ключевые слова: ожидаемая продолжительность жизни, машинное обучение, модели машинного обучения.

Благодарности. Работа выполнена при поддержке Программы стратегического академического лидерства РУДН.

Для цитирования. Бру Куамэ Амос, И. Смирнов. Детерминирующие факторы в прогнозировании ожидаемой продолжительности жизни с помощью машинного обучения. Advanced Engineering Research. — 2022. — Т. 22, № 4. — С. 373–383. https://doi.org/10.23947/2687-1653-2022-22-4-373-383

Introduction

Human life expectancy can be understood as a statistic used in demography to estimate the average age at which people in a given region and at a given time can be expected to live under current conditions [1]. Life expectancy is not only a statistical indicator of human health, but also a means of assessing the degree of economic, educational, health [2] and environmental development. It should be noted that the World Health Organization (WHO) considers life expectancy to be a key, if not the most important, indicator of health, reflecting the instruments of human existence [3–6]. In the majority of the world countries, life expectancy has in fact increased. Global life expectancy increased from 67.2 years in 2005 to 70.8 years in 2015[1]. The United Nations and individual national governments now have the optimization of human life expectancy, health and well-being as their main objective [7–8]. The UN has been a strong promoter of human health by providing sanitary remedies, which greatly improves the urban environment and helps developing countries[2]. In the African region of the World Health Organization, life expectancy is 61.2 years, while in the European region, it is 77.5 years, giving a ratio of 1.3 between the two regions [9]. Analysis of the disparities in life expectancy between developed and developing countries will enable the United Nations to improve its health promotion and humanitarian assistance activities. It will also enable governments of different nations to establish more effective policies to increase life expectancy and improve living standards. States would be able to significantly increase the life expectancy of their population by investing more in the health care system [7]. According to [10], the increase in life expectancy in the United States as a function of per capita income is substantially related to the increase in income level. According to some researchers, the relevant factors affecting life expectancy are mainly environmental, social and economic factors. These vary according to location and involve economic development, medical and health requirements. Existing research has

[1] United Nations Statistical Yearbook, 2017 edition. United Nations, New York; 2017.
[2] United Nations Economic and Development website: https://www.un.org/chinese/esa/health.htm (accessed: 9 February, 2021)

given rise to debates about the factor that determines life expectancy. In [11], the author analysed life expectancy in Tibet, China, and found the main factors determining life expectancy. The author in [12], thought that social economy played an important role in determining life expectancy in the early stages of development. However, it was replaced by diet and lifestyle when economic development reached a certain level. The article [13] considered that the determining factor in the evolution of life expectancy in Eastern Europe was lifestyle.

In the end, several studies have identified numerous factors influencing life expectancy. However, few studies compared economic development with environmental factors to analyse the intensity of their impact on life expectancy. Several studies are trying to find out the determining factor of life expectancy. It must be said that there are several. Epidemiological studies in developed countries reveal large differences in life expectancy that are often highly complex. A current study in the United States suggests that 10 to 38 % of the differences in life expectancy can be explained by work-related stress. Life expectancy depends on many factors such as economic status, regional changes in education, gender disparities, physical and mental illnesses, alcohol consumption, GDP, health care spending, and many other demographic factors. Life expectancy has actually increased during the 20th and 21st centuries in industrialized countries [14–20]. The improvement in life expectancy in Europe is followed by a population growth in the over-50 age group. There were 179 million people aged 50 or over in all EU Member States in 2008, and 195 million in 2014, with women accounting for about 55 % of the total [21–24].

According to some sources, the level of economic development has a significant effect on life expectancy. Indeed, studies have shown that people who are financially well off and those from wealthy families tend to have a higher life expectancy [25, 26]. For some researchers, other economic development variables such as GDP per capita [27, 28], urbanisation rate [29] and level may affect life expectancy to different degrees. Some studies show that environmental factors are determinants of life expectancy [29]. Indeed, according to [29, 30], most environmental factors, such as ecological resilience and environmental sustainability, are positively correlated with life expectancy, while some factors, including biodiversity, are negatively correlated with life expectancy. In [31], current environmental conditions influence the life expectancy of the population at birth, while cumulative changes in circumstances continue to influence the remaining life expectancy of the population at different ages over time. J. O. Anderson in [32] thought that people living in an environment with high levels of particulate pollutants over a long period of time had higher cardiovascular morbidity, and that there was some degree of dose dependence. Other researchers have studied the impact of different environmental variables on life expectancy. A. Wuffle [33] compared the average temperature of all US states. The results showed that the lower the average temperature in November, the higher the life expectancy of the population in those states [34–36].

Therefore, through machine learning, we will determine the factors that influence life expectancy. Machine learning (ML) can be understood as a discipline that lies at the intersection of mathematics, statistics and computer science. Machine learning has played an important role in the development of artificial intelligence (AI). Thus, artificial intelligence, through machine learning, helps companies to prevent problems and increase profits. In the field of health, machine learning is still surprising researchers. It is now the most widely used tool for prediction and forecasting. Machine learning, which represents a cutting-edge technology due to its predictive accuracy in several problems, is widely used to increase life expectancy by reducing the mortality rate [17]. Indeed, given that several elements impact on life expectancy, the multiple regression model is of paramount importance and corresponds to the exploration of the specific relationship and level of impact between several factors and life expectancy.

This paper uses multiple linear regression models and Pearson's correlation coefficient to examine the relationship between several variables on life expectancy and provide more help for future research on both sides. These models are also used as a basis for suggestions to states for improving life expectancy in order to achieve a development of human society.

Materials and Methods

The World Health Organization (WHO) Global Health Observatory (GHO) data repository tracks health status and many other related factors for all countries. The datasets are made available to the public for analysis of health data. Data on life expectancy and health factors for 193 countries were collected on the same WHO website and the corresponding economic data were collected on the UN website. From all categories of health-related factors, only the most representative critical factors were selected. It has been observed that in the last 15 years, the health sector has undergone enormous development, resulting in improved human mortality rates, especially in developing countries, compared to the last 30 years. Therefore, in this project, we considered data from the year 2000 to 2015 for 193 countries for further analysis. The individual data files were merged into a single dataset. Initial visual inspection of the data revealed some missing values. As the datasets were from WHO, we did not find any obvious errors.

Our dataset had missing data, and the missing data were for population, hepatitis B and GDP. Imputation of missing data using the *mice* function in the R package of the same name.

Each variable is associated with an imputation model, conditional on the other variables in the data set: if we have X_k variables, the missing data for the variable X_i will be replaced by the predictions of a model created from the other variables.

The final file is composed of 22 attributes, the target variable "life expectancy" and various other social factors, such as total expenditure on life, population, education and health factors, such as BMI, measles, etc. These data are available on Kaggle [18]. All predictors were then divided into several broad categories: vaccination-related factors, mortality factors, economic factors, and social factors.

Data mining

The objective of this section is to gain a better understanding of the data by extracting information from the data. We mainly want to determine the relationship between the variables.

The correlation matrix, visualized using a heat map (as shown in Figure 1), is one of the best ways to understand the correlation between variables. It is plotted using the R library "Reshape2" and shows us the strength of the linear relationships between the variables. The linear relationship between the outcome and the characteristics can be estimated by a correlation matrix. In multivariate analysis, it plays an important role, as it elaborates the relationship between the different components [19]. Looking at Figure 1, we can see that:

the target variable Life expectancy is strongly correlated (positively or negatively) with:

– Adult mortality (-0.70);

– HIV/AIDS (-0.56);

– Composition of income resources (0.72);

– Schooling (0.75).

There is also a very low correlation between the target variable Life Expectancy and Population (-0.02) or no correlation at all.

The child deaths variable is extremely positively correlated with deaths under five years of age (1.00).

The GDP variable and the percentage of expenditure are positively correlated (0.90).

The Hepatitis B variable is moderately positively correlated with Polio and Diphtheria (0.49) and (0.61).

The variables diphtheria and polio are strongly positively correlated (0.67).

The HIV/AIDS variable is negatively correlated with resource composition (-0.25).

The thinness variable 10 ... 19 years is very strongly positively correlated with the thinness variable 5 ... 9 years (0.94).

The variable Schooling and the income composition of resources are very strongly correlated (0.8).

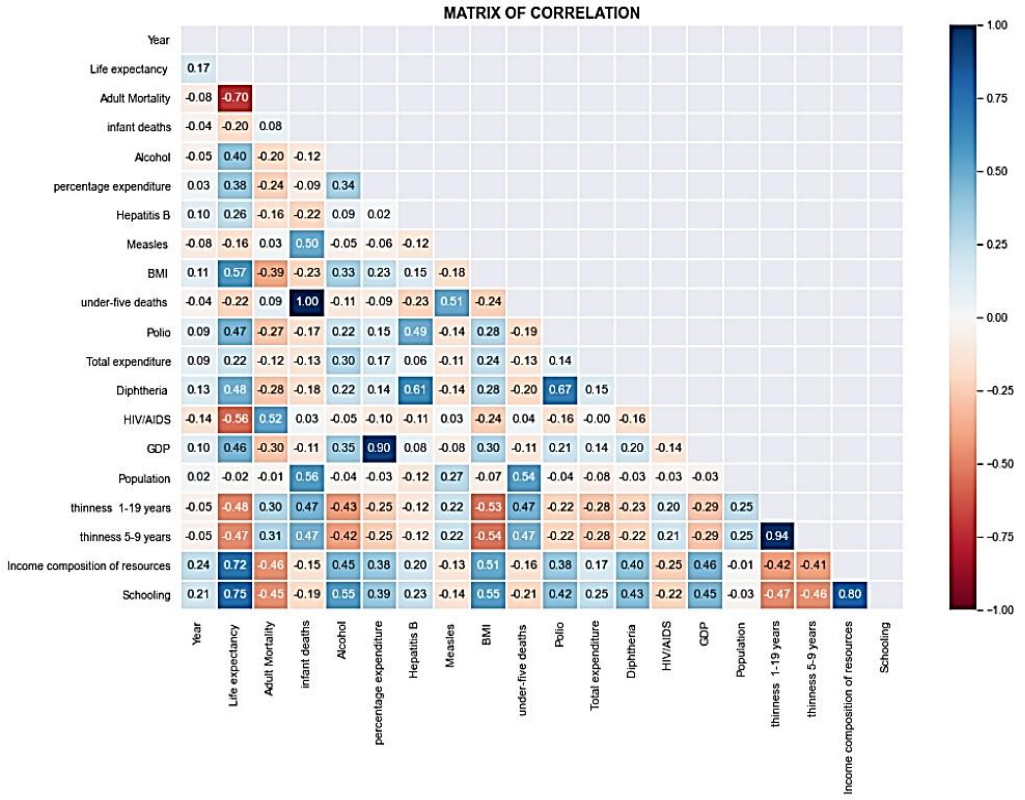

Fig. 1. Visualisation of the correlation matrix by heat map (figure by authors)

By examining the correlation coefficients in Figure 1, we detect potential predictors of life expectancy. For each numerical variable that is potentially predictive of life expectancy, we will run a simple linear regression between it and the life expectancy variable, display the Pearson correlation coefficient and its P
y = a$x + b$. y is the dependent variable and x is the independent variable. a and b are the model parameters (a is the slope of the fitted line and b is the intercept).

From the result of this exploratory analysis on our data, we concluded that adult mortality, HIV/AIDS, BMI, income composition, and education are the most important factors in predicting life expectancy. This selection was made on the basis of the Pearson correlation coefficient and p-value (as shown in Figures 2, 3, 4, 5, 6).

– The correlation between the variable infant mortality, GDP, alcohol, percentage of expenditure, hepatitis B, measles, under-five deaths, polio, total expenditure, diphtheria, population, age 1–19, age 5–9, and life expectancy is statistically significant as their p-values are less than 0.001, but the linear relationship between these variables is weak with a Pearson correlation coefficient of less than 0.5. Under these conditions, the variables child deaths, GDP, alcohol, percentage of expenditure, hepatitis B, measles, under-five deaths, polio, total expenditure, diphtheria, population, leanness 1–19 years, leanness 5–9 years, cannot be considered as a predictor of life expectancy.

– There is a strong negative correlation between the variables Adult Mortality and Life Expectancy with a Pearson correlation coefficient of –0.7 and statistically significant since the p-value is less than 0.001. In other words, as adult mortality increases, life expectancy decreases. Under these conditions, the adult mortality variable can be considered a predictor of life expectancy.

– There is a negative correlation between the variables HIV/AIDS and life expectancy with a Pearson correlation coefficient of -0.56 and statistically significant since the p-value is less than 0.001. As the number of people affected by HIV/AIDS increases, life expectancy decreases. Under these conditions, the HIV/AIDS variable can be considered a predictor of life expectancy.

– There is a positive correlation between the BMI and life expectancy variables with a Pearson correlation coefficient of 0.56 and statistical significance since the p-value is less than 0.001. Under these conditions, the BMI variable can be considered a predictor of life expectancy.

– There is a strong positive correlation between the variables Income composition of resources and life expectancy with the Pearson correlation coefficient of 0.69 and statistically significant as the p-value is less than 0.001. The graph shows that as the composition of income increases, life expectancy increases. Under these conditions, the variable Income composition of resources can be considered as a predictor of life expectancy.

– There is a strong positive correlation between the variables Education and Life Expectancy with a Pearson correlation coefficient of 0.72 and statistically significant as the p-value is less than 0.001. The graph shows that the higher the education, the higher the life expectancy. Under these conditions, the education variable can be considered a predictor of life expectancy.

Fig. 2. Correlation between life expectancy and HIV/AIDS variable (figure by authors)

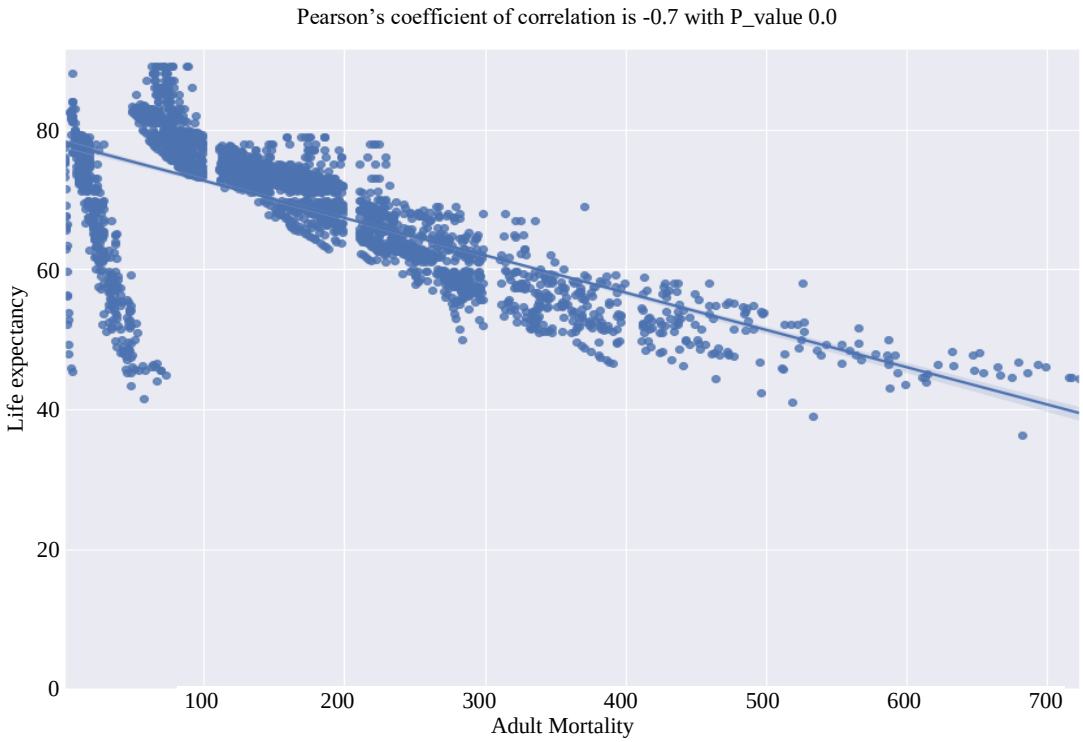

Fig. 3. Correlation between life expectancy and the adult mortality variable (figure by authors)

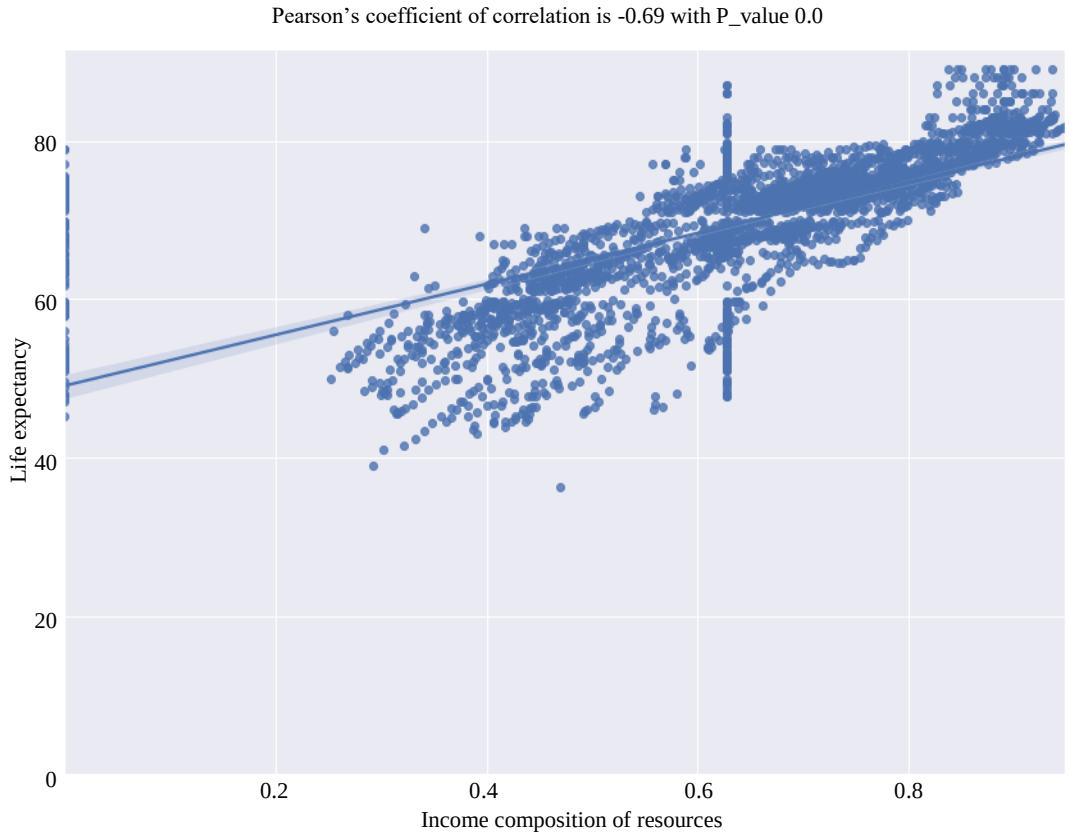

Fig. 4. Correlation between life expectancy and the income composition variable (figure by authors)

Fig. 5. Correlation between life expectancy and the BMI variable (figure by authors)

Fig. 6. Correlation between life expectancy and the education variable (figure by authors)

Based on the results of the Pearson correlation, we detect the variables that have an influence on life expectancy. However, the Pearson correlation is not sufficient to determine the predictors of life expectancy. For this purpose, we will run several regression models to select the one with the smallest AIC, the highest fit, and the smallest mean square error, R^2 and the smallest root mean square error (RMSE).

Methodology

In order to determine the variables that predict life expectancy, different regression models are used, namely, multiple linear regression, rigid regression and lasso regression. We will then examine the criteria for selecting (p – 1) explanatory variables from the k available explanatory variables. These criteria are: Mallows' Cp criterion, the coefficient of determination, R^2 The Bayesian information criterion (BIC), the Akaike information criterion (AIC).

Multiple linear regression

Multiple linear regression is an immediate generalization of simple linear regression. In multiple linear regression, the function F that we want to estimate no longer depends on a single variable, but on several. If we have n pairs of the form (X_i = ($x_{i,1}$, $x_{i,2}$, ..., $x_{i,m}$) ϵR^m, $y_i \epsilon R^m$), with y_i the result obtained for the observation X_i = ($x_{i,1}$, $x_{i,2}$, ..., $x_{i,m}$), then the function we wish to estimate will be of the general form below :

$$F\ (X_i = (x_{i,1}, x_{i,2}, ..., x_{i,m}) = a_1 x_{i,1} + a_2 x_{i,2} + ...+ a_m x_{i,m} + a_0. \quad (1)$$

The objective is to estimate the vector A = a_1, a_2 , ..., a_m, a_0 so that the function F is as close as possible to y_i .

As with simple linear regression, the least squares method can be used to find the vector A, and the function to be minimized will be defined as follows:

$$E = \sum_{i=1}^{n} [\ y_i - F\ (x_{i,1}, x_{i,2}, ..., x_{i,m})]^2. \quad (2)$$

One of the most difficult aspects of abundant regression algorithms is to determine how to converge to the configurations (a_1, a_2 , ..., a_m, a_0) that yield errors $\mathcal{E} = \hat{y}_i - y_i$ and avoid the trap of over-learning.

The best-known approach to minimizing the error calculation function E while avoiding over-learning is the introduction of the concept of regularization.

There are two regularizations widely used with regression models: the Lasso regularization and the Ridge regularization.

Lasso and Ridge regression

Lasso regression is a regression model in which the selection and regulation of variables take place simultaneously. This method uses a penalty that affects the value of the regression coefficients. With Lasso regularization, the error function to be minimized becomes:

$$E_{Lasso} = \sum_{i=1}^{n} [\ y_i - F\ (x_{i,1}, x_{i,2}, ..., x_{i,m})]^2 + \lambda \sum_{i=0}^{m} |\ a_i|. \quad (3)$$

The difference between E and E_{Lasso} is that in E_{Lasso} we have added the term $\lambda \sum_{i=0}^{m} |\ a_i|^2$ to further sanction solutions with values of (a_1, a_2 , ..., a_m, a_0).

Ridge regression is a regularised regression algorithm that performs an L2 regularisation by adding an L2 penalty, which is equal to the square of the magnitude of the coefficients. With Ridge regularization, large values of (a_1, a_2 , ..., a_m, a_0) are more protected, and the error function to be minimized becomes:

$$E_{Ridge} = \sum_{i=1}^{n} [\ y_i - F\ (x_{i,1}, x_{i,2}, ..., x_{i,m})]^2 + \lambda \sum_{i}^{m} a_i^2. \quad (4)$$

We note that in both the Lasso and Ridge regularization cases, when the value λ is set to 0, then E = E_{Lasso} = E_{Ridge}.

Mallows' Cp criterion

Mallows' Cp is a selection criterion between several regression models. It compares the accuracy and bias of the full model with those of models containing a subset of predictors. Mallows' Cp criterion is defined from the following formula:

$$C_p = \frac{SC_{res}}{\widehat{\delta^2}} - (n - 2p). \quad (5)$$

But the problem is that we can no longer estimate δ^2 by $s^2 = \frac{SC_{res}}{n-p}$ because C_p would always be equal to p and then it would no longer be interesting [9]. So, in practice, we estimate δ^2 by the s^2 of the model that involves all k explanatory variables of the available model, then we choose among the models the one for which Mallows' Cp criterion is closest to p.

The coefficient of determination R^2

The website R^2 is the simplest to use. However, with the introduction of new variables, it increases monotonically even if they are weakly correlated with the explained variable. It is therefore advisable to turn to the use of other criteria such as the adjusted R^2adjusted criterion, Mallows' Cp, the AIC and AICc criteria, the BIC criterion.

The adjusted coefficient of determination R^2 is the evolved version of the coefficient of determination R^2.

The adjusted R^2 determines the amount of variance of the dependent variable, which can be explained by the independent variable. On the basis of the fitted value R^2 value, one can judge whether the data in the regression equation are appropriate. The higher the R^2 the higher the fitted value, the better the regression equation because it implies that the independent variable is chosen to determine the dependent variable.

The Bayesian information criterion BIC

The Bayesian Information Criterion (BIC) is derived from the Akaike Information Criterion (AIC) and is defined by : BIC = -2 log (L) + k log (n). It is more parsimonious than the AIC criterion because it penalises the number of variables present in the model more. According to [9], the AIC was introduced to retain the variables relevant to the forecast, whereas the BIC criterion aims at selecting the statistically significant variables in the model.

The Akaike Information Criterion (AIC)

The Akaike Information Criterion (AIC) is a mathematical method that is applied to models estimated by a maximum likelihood method to assess how well a model fits the data from which it was generated. AIC is applied to analysis of variance, multiple linear regression, logistic regression and Poisson regression. The AIC criterion is defined by :

$$AIC = 2K - 2Log(L), \quad (6)$$

where L is the maximised likelihood and k is the number of model parameters. With this criterion, the deviance of the model -2 log (L) is penalised by two times the number of parameters.

Therefore, the AIC represents a compromise between bias, which decreases with the number of parameters, and parsimony, the desire to describe the data with as few parameters as possible.

The rigour would dictate that all models compared should derive from the same 'complete' model included in the list of models compared.

The best model is the one with the lowest AIC. When the number of parameters k is large compared to the number of observations n, i.e., if N/k < 40, it is recommended to use the corrected AIC. The corrected Akaike information criterion, AICc, is defined by:

$$AICc = AIC + \frac{2K(K+1)}{n-K-1}. \quad (7)$$

Results

Our analysis has shown that life expectancy increases over the years, and that it is on average higher in developed countries than in developing countries.

This study has also led us to the conclusion that the model chosen for the selection of life expectancy predictors is multiple linear regression (Table 1), as this model obtained the lowest Akaike information criterion of 6109.07, an adjusted coefficient of 85% and an RMSE of 3.85. R^2 of 85 % and an RMSE of 3.85.

These measures were better than those of the Lasso and Ridge regression models. According to this model and following the p-value of less than $2.2e^{-16}$, all variables are significant, except for: Alcohol, Hepatitis B, Measles, Population, Slimness.1.19, Slimness.5.9. This means that we can do without these variables to explain life expectancy. However, applying the Akaike information criterion to the multiple linear regression model, the variables that best explain life expectancy are: adult mortality, infant deaths, percentage of expenditure, measles, under-five deaths, polio, total expenditure, diphtheria, HIV/AIDS, GDP, thinness. 1.19 years, income composition, and school enrolment.

Table 1

Models	Adjusted R^2	RMSE
Multiple Linear Regression	**0.85**	**3.85**
Lasso Regression	0.82	3.85
Ridge Regression	0.82	3.91

Conclusion

Before analysing this data set, we had the impression that life expectancy could be increased if we had more money. This is because it takes money to be healthy and to receive appropriate medical treatment. Moreover, if a country is economically developed (GDP), all its citizens can afford appropriate medical treatment. This would mean that life expectancy depends largely on economic factors. However, after analysing this dataset, we have concluded that life expectancy is mainly affected by adult mortality, infant mortality, percentage of expenditure, measles, under-five mortality, polio, total expenditure, diphtheria, HIV/AIDS, GDP, wasting. 1.19 years, income composition, and schooling. This makes sense because if a person is educated enough to recognize health problems, they can make appropriate lifestyle changes, including but not limited to diet and exercise, which would ideally extend their life expectancy. Education can change a person's perception and help them understand the benefits of being fit and its impact on health. In addition, a higher level of education could be linked to a higher income, and a higher income would mean higher spending on health and fitness. Thus, education is directly or indirectly a good predictor of life expectancy. Various machine learning models have been used for training. Among these models, the multiple linear model has proven to be very effective in determining both the coefficient of determination and the errors. This model can therefore be used for the prediction of life expectancy.

References

1. Arias E. United States Life Tables, 2009. National Vital Statistics Reports. 2014; 62:1–63.

2. Yafei Wu, Ke Hu, Yaofeng Han, et al. Spatial Characteristics of Life Expectancy and Geographical Detection of Its Influencing Factors in China. International Journal of Environmental Research and Public Health. 2020;17:906. https://doi.org/10.3390/ijerph17030906

3. Ming Wen, Danan Gu. Air Pollution Shortens Life Expectancy and Health Expectancy for Older Adults: The Case of China. The Journals of Gerontology: Series A. 2012;67:1219–1229. https://doi.org/10.1093/gerona/gls094

4. Cervantes PAM, López NR, Rambaud SC. The Relative Importance of Globalization and Public Expenditure on Life Expectancy in Europe: An Approach Based on MARS Methodology. International Journal of Environmental Research and Public Health. 2020;17:8614. http://dx.doi.org/10.3390/ijerph17228614

5. Reynolds MM, Avendano M. Social Policy Expenditures and Life Expectancy in High-Income Countries. American Journal of Preventive Medicine. 2018;54:72–79. https://doi.org/10.1016/j.amepre.2017.09.001

6. Sede IP, Ohemeng W. Socio-economic determinants of life expectancy in Nigeria (1980–2011). Health Economics Review. 2015;5:1–11.

7. Daquan Huang, Shuimiao Yang, Tao Liu. Life Expectancy in Chinese Cities: Spatially Varied Role of Socioeconomic Development, Population Structure, and Natural Conditions. International Journal of Environmental Research and Public Health. 2020;17:6597. http://dx.doi.org/10.3390/ijerph17186597

8. Okamoto, K. Life Expectancy at Age 65 and Environmental Factors: An Ecological Study in Japan. Archives of Gerontology and Geriatrics. 2006;43:85–91. http://dx.doi.org/10.1016/j.archger.2005.09.005

9. WHO. Ambient Air Pollution: A Global Assessment of Exposure and Burden of Disease. World Health Organization, 2016. 121 p. https://apps.who.int/iris/handle/10665/250141

10. Xinjie Zha, Yuan Tian, Xing Gao, et al. Quantitatively Evaluate the Environmental Impact Factors of the Life Expectancy in Tibet, China. Environmental Geochemistry and Health. 2019;41:1507–1520. https://link.springer.com/article/10.1007/s10653-018-0211-z

11. Nkalu CN, Edeme RK. Environmental Hazards and Life Expectancy in Africa: Evidence from GARCH Model. *SAGE* Open; 2019, 9. https://doi.org/10.1177/2158244019830500

12. Inglehart Ronald, Christian Welzel. How Development Leads to Democracy What We Know About Modernization. Foreign Affairs. 2009;88:33–48.

13. Cockerham WC. The Social Determinants of the Decline of Life Expectancy in Russia and Eastern Europe: A Lifestyle Explanation. Journal of Health and Social Behavior. 1997;38:117–130.

14. Jessica Y Ho, Arun S Hendi. Recent Trends in Life Expectancy across High Income Countries: Retrospective Observational Study. BMJ. 2018;362:k2562. https://doi.org/10.1136/bmj.k2562

15. Penuelas J, Krisztin T, Obersteiner M, et al. Country-Level Relationships of the Human Intake of N and P, Animal and Vegetable Food, and Alcoholic Beverages with Cancer and Life Expectancy. International Journal of Environmental Research and Public Health. 2020;17:7240. https://doi.org/10.3390/ijerph17197240

16. Sidey-Gibbons JAM, Sidey-Gibbons CJ. Machine Learning in Medicine: A Practical Introduction. BMC Medical Research Methodology. 2019;19:1–18. https://bmcmedresmethodol.biomedcentral.com/ articles/10.1186/s12874-019-0681-4

17. Malpe V, Tugaonkar P. Machine Learning Trends in Medical Sciences. In: Proc. 2nd International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud), August 2018. P. 495–499.

18. KumarRajarshi. WHO. Life Expectancy. Statistical Analysis on Factors Influencing Life Expectancy. https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who

19. Thu Pham-Gia, Vartan Choulakian. Distribution of the Sample Correlation Matrix and Applications. Open Journal of Statistics. 2014;4:48571. 10.4236/ojs.2014.45033

20. Svensson K. Predicting Life Expectancy Using Machine Learning. 2018. https://www.semanticscholar.org/paper/Predicting-Life-Expectancy-Using-Machine-Learning-Svensson/984adbb5aee16d38a6686895dda2afd3087b2261

21. Müller AC, Guido S. Introduction to Machine Learning with Python.,1st ed. O'Reilly Media, Inc.; 2016.

22. Ki-Young Lee, Kyu-Ho Kom, Jeong-Jin Kang, et al. Comparison and Analysis of Linear Regression and Artificial Neural Network. International Journal of Applied Engineering Research. 2017;12:9820–9825.

23. Paranjape RS, Challacombe SJ. HIV/AIDS in India: An Overview of the Indian Epidemic. Oral Diseases. 2016;22:10–14. http://dx.doi.org/10.1111/odi.12457

24. Haebong Woo. Patterns and Evolution of Life Span Inequality Using the Gini Coefficient. Pogŏn Sahoe Yŏn'gu. 2013;33:419–451. 10.15709/hswr.2013.33.4.419. https://www.researchgate.net/publication/275246163_Patterns_and_Evolution_of_Life_Span_Inequality_Using_the_Gini_Coefficient

25. Chen Wu. Human Capital, Life Expectancy, and the Environment. Journal of International Trade and Economic Development. 2017;26:885–906.

26. Carolina Cosculluela-Martínez, Raquel Ibar Alonso, Geoffrey JD Hewings. Life Expectancy Index: Age Structure of the Population and Environmental Change. Social Indicators Research. 2019;142:507–522.

27. Muhamad Haroon Shah, Nianyong Wang, Irfan Ullah, et al. Does Environment Quality and Public Spending on Environment Promote Life Expectancy in China? Evidence from a Nonlinear Autoregressive Distributed Lag Approach. International Journal of Health Planning and Management. 2021;36:545–560. http://dx.doi.org/10.1002/hpm.3100

28. Mariani F, Pérez-Barahona A, Raffin, N. Life Expectancy and the Environment. Journal of Economic Dynamics and Control. 2010;34:798–815.

29. Tuljapurkar Sh, Horvitz CC. From Stage to Age in Variable Environments: Life Expectancy and Survivorship. Ecology. 2006;87:1497–1509. http://dx.doi.org/10.1890/0012-9658(2006)87[1497:FSTAIV]2.0.CO;2

30. Kampa M, Castanas E. Human Health Effects of Air Pollution. Environmental Pollution. 2008;151:362–367. http://dx.doi.org/10.1016/j.envpol.2007.06.012

31. Tagaris E, Kuo-Jen Liao, DeLucia AJ, et al. Potential Impact of Climate Change on Air Pollution-Related Human Health Effects. Environmental Science and Technology. 2009;43:4979–4988.

32. Anderson JO, Thundiyil JG, Stolbach, A. Clearing the Air: A Review of the Effects of Particulate Matter Air Pollution on Human Health. Journal of Medical Toxicology. 2012;8:166–175. http://dx.doi.org/10.1007/s13181-011-0203-1

33. Wuffle A, Brians CL, Coulter K. Taking the Temperature: Implications for the Adoption of Election Day Registration, State level Voter Turnout, and Life Expectancy. PS: Political Science & Politics**.** 2012;45:78–82.

34. Brunner E, Maruyama K. SP4-32 Health and Sustainability: An International Ecological Study of Carbon Dioxide Emissions and Life Expectancy. Journal of Epidemiology and Community Health. 2011;65:A442–A443.

35. Clootens N. Public Debt, Life Expectancy, and the Environment. Environmental Modeling & Assessment. 2017;22:267–278.

36. Tetzlaff F, Epping J, Sperlich S, et al. Widening Income Inequalities in Life Expectancy? Time Trend Analysis Using German Health Insurance Data. Journal of Epidemiology and Community Health. 2020;74:592–597. 10.1136/jech-2019-212966

Received 14.09.2022.
Revised 17.10.2022.
Accepted 18.10.2022.

About the Authors:

Brou Kouame Amos, postgraduate of the Information Technology Department, Peoples' Friendship University of Russia (RUDN) (6, Miklikho-Maklaya St., Moscow, 117198, RF), broukouameamos9@gmail.com

Smirnov, Ivan V., associate professor of the Information Technology Department, Peoples' Friendship University of Russia (RUDN) (6, Miklikho-Maklaya St., Moscow, 117198, RF), Cand.Sci. (Phys.-Math.), associate professor, ORCID, Scopus, smirnov-iv@rudn.ru

Claimed contributorship:

Brou Kouame Amos: basic concept formulation; research objectives and tasks; data pre-processing; analysis of research results. I. V. Smirnov: work control; the text revision; correction of the conclusions.

Conflict of interest statement

The authors do not have any conflict of interest.

All authors have read and approved the final manuscript.

Об авторах:

Бру Куамэ Амос, аспирант кафедры «Информационные технологии» Российского университета дружбы народов (117198, РФ, г. Москва, ул. Миклухо-Маклая, 6), broureino9@gmail.com

Смирнов Иван Валентинович, доцент кафедры «Информационные технологии», Российского университета дружбы народов» (117198, РФ, г. Москва, Миклухо-Маклая, 6), кандидат физико-математических наук, доцент, Scopus, ORCID, smirnov-iv@rudn.ru

Заявленный вклад соавторов:

К. А. Бру: формирование основной концепции, цели и задачи исследования, предварительная обработка данных и анализ результатов исследований, подготовка текста, формирование выводов, сбор данных и доработка текста. И. В. Смирнов: контроль за работой, доработка текста и корректировка выводов.

Конфликт интересов

Авторы заявляют об отсутствии конфликта интересов.

Все авторы прочитали и одобрили окончательный вариант рукописи.